Bioinformatics

CS300

Genome annotation and sequence-based gene prediction

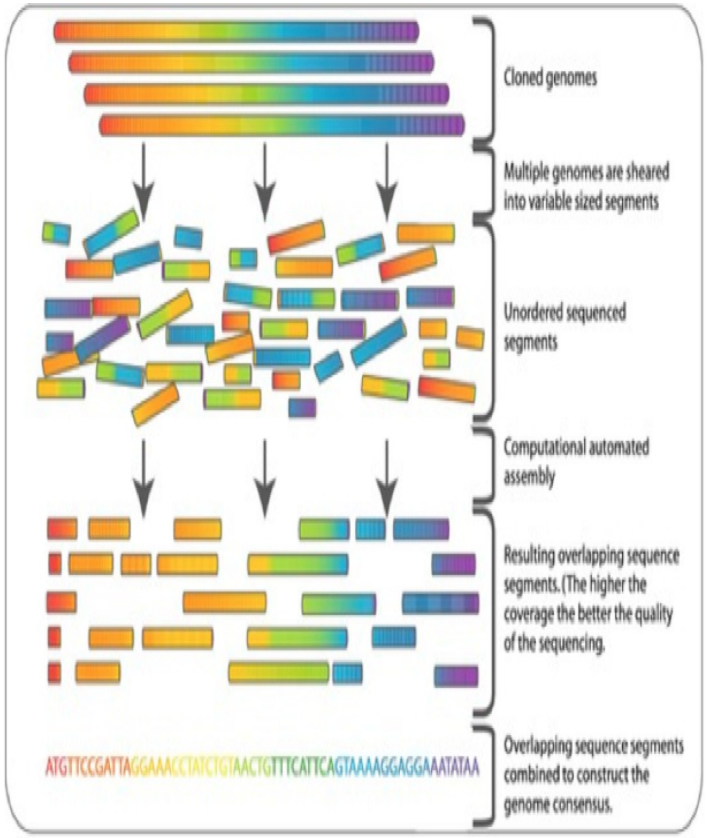
Week10, Deck 2

Fall 2022

Oliver BONHAM-CARTER

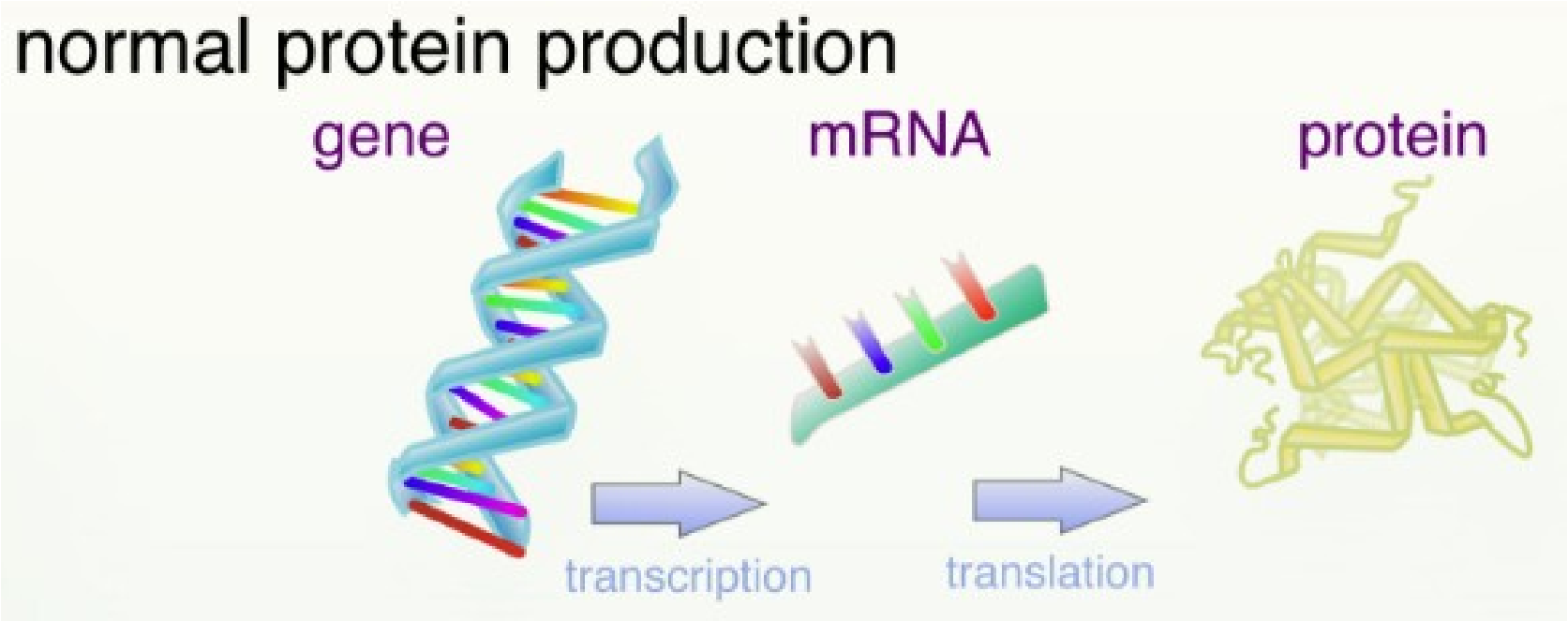
Genome Projects

- **Goals**:
  - Determine complete genome sequence of an organism
  - **Annotate protein-coding genes and other important genome-encoded features**
    - find
    - identify
    - characterize
    - describe
    - computational predictions later confirmed at the lab bench



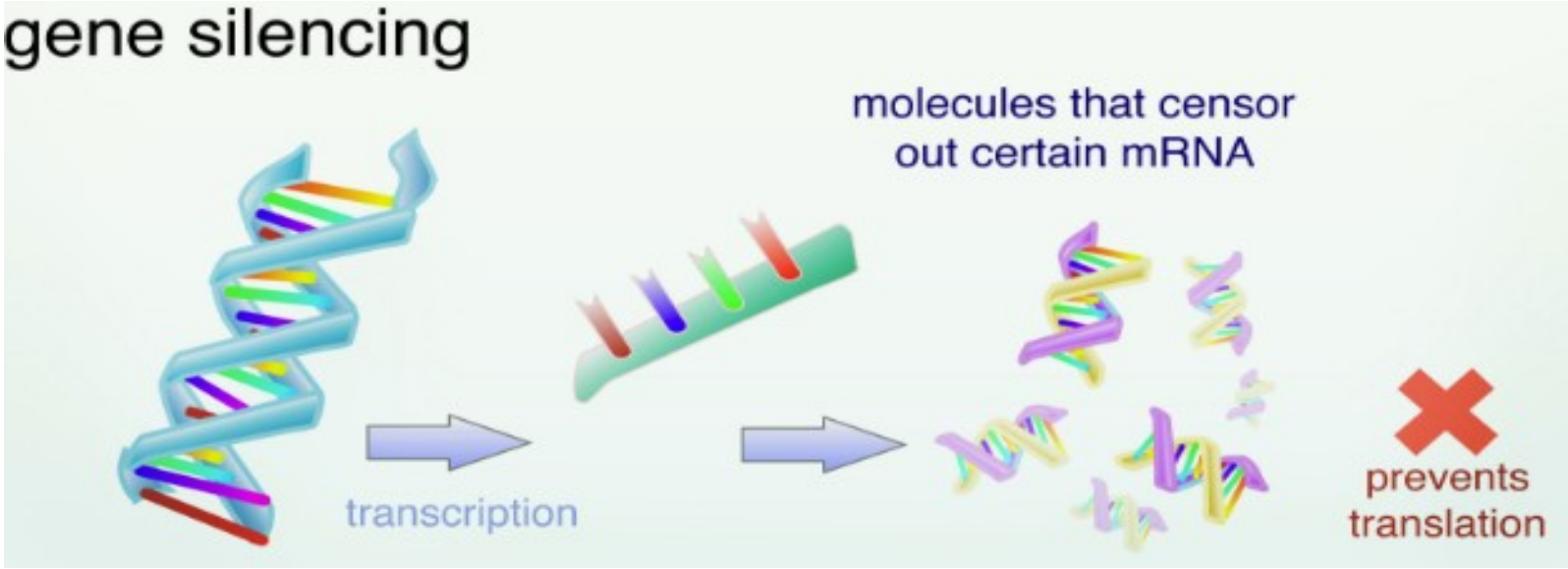
Gene Prediction

- Sequence-based – find features based on specific sequences
- What does a gene look like?
  - Qualities?
  - Behaviors?
  - Sequence trends?



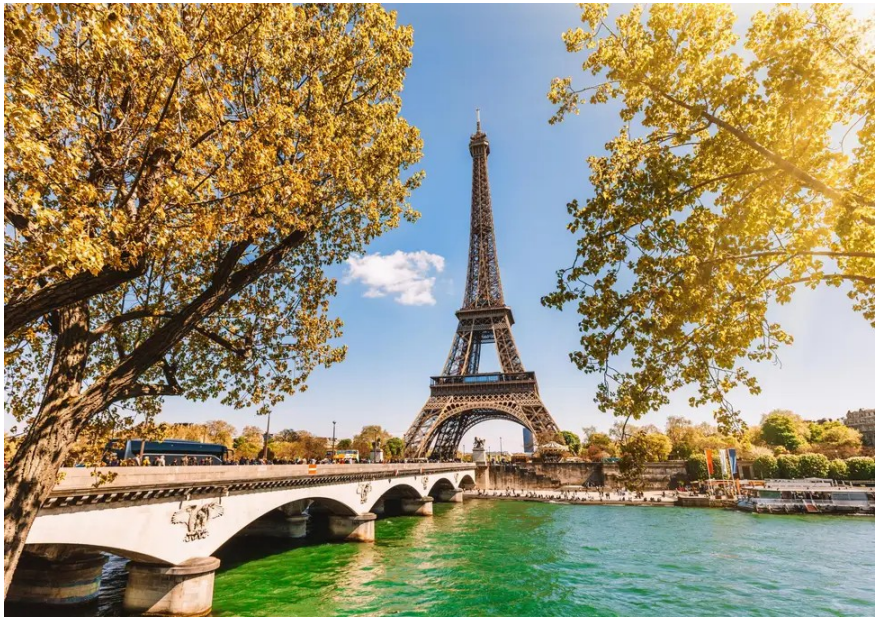
Gene Prediction

- Two obvious questions:
- **Which proteins are available in the genome?**
- **Would proteins inform us about present genes?**



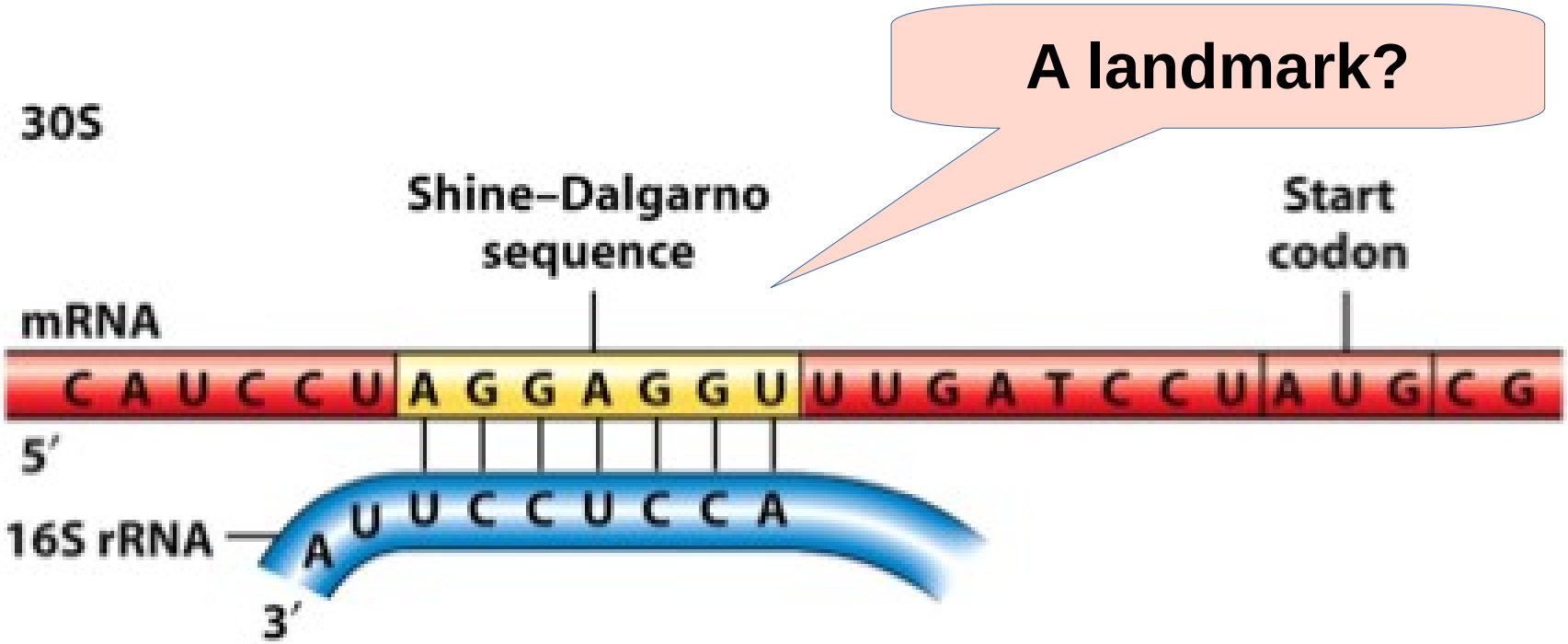
Not all present genes will make a protein ...

What are *Land Marks*?

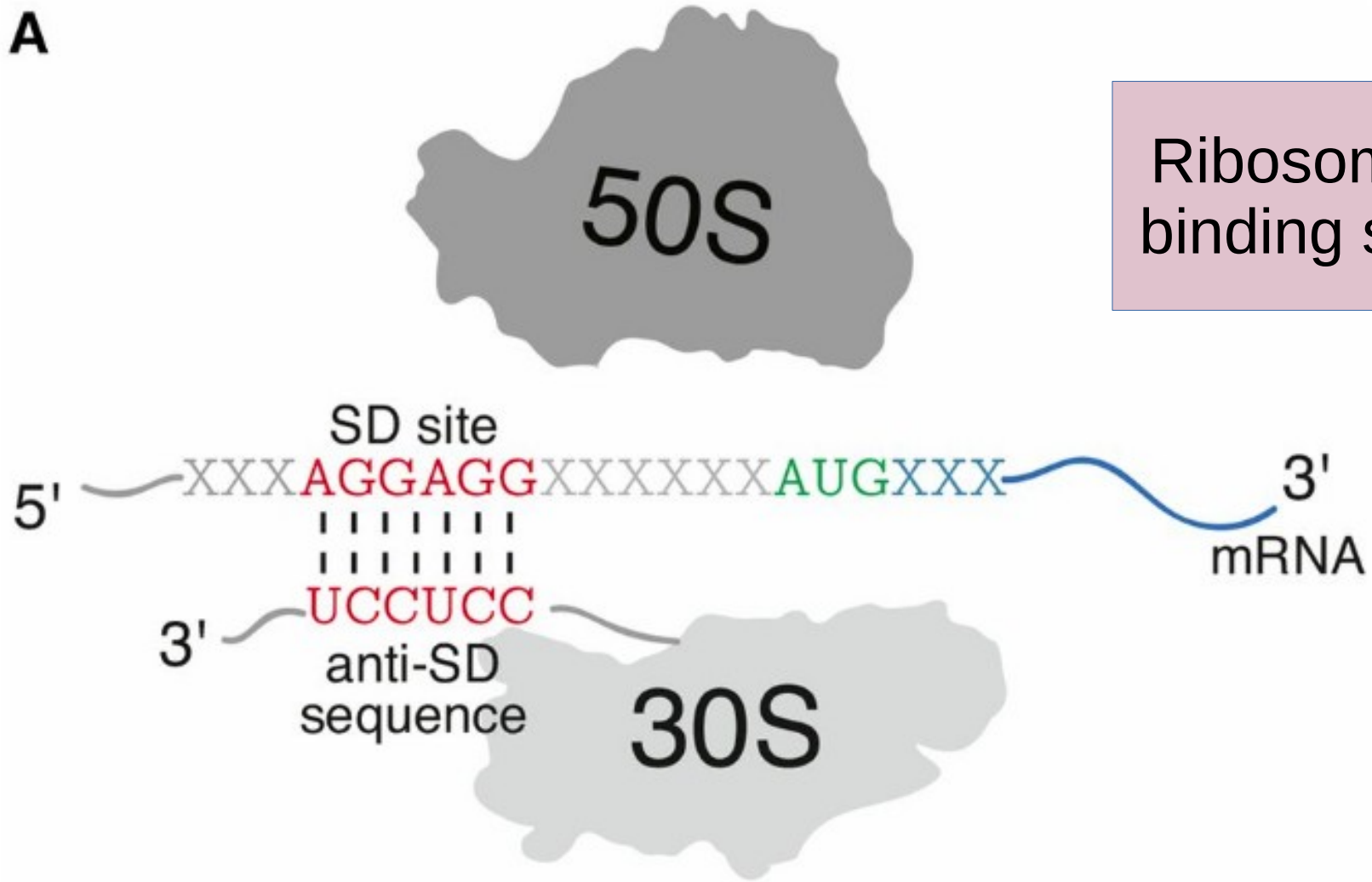


Shine-Dalgarno Sequence

- The Shine–Dalgarno (SD) sequence is a ribosomal binding site in bacterial and archaeal messenger RNA, generally located around 8 bases upstream of the start codon AUG
- The RNA sequence helps recruit the ribosome to the messenger RNA (mRNA) to initiate protein synthesis by aligning the ribosome with the start codon.



Genetic Land Marks?



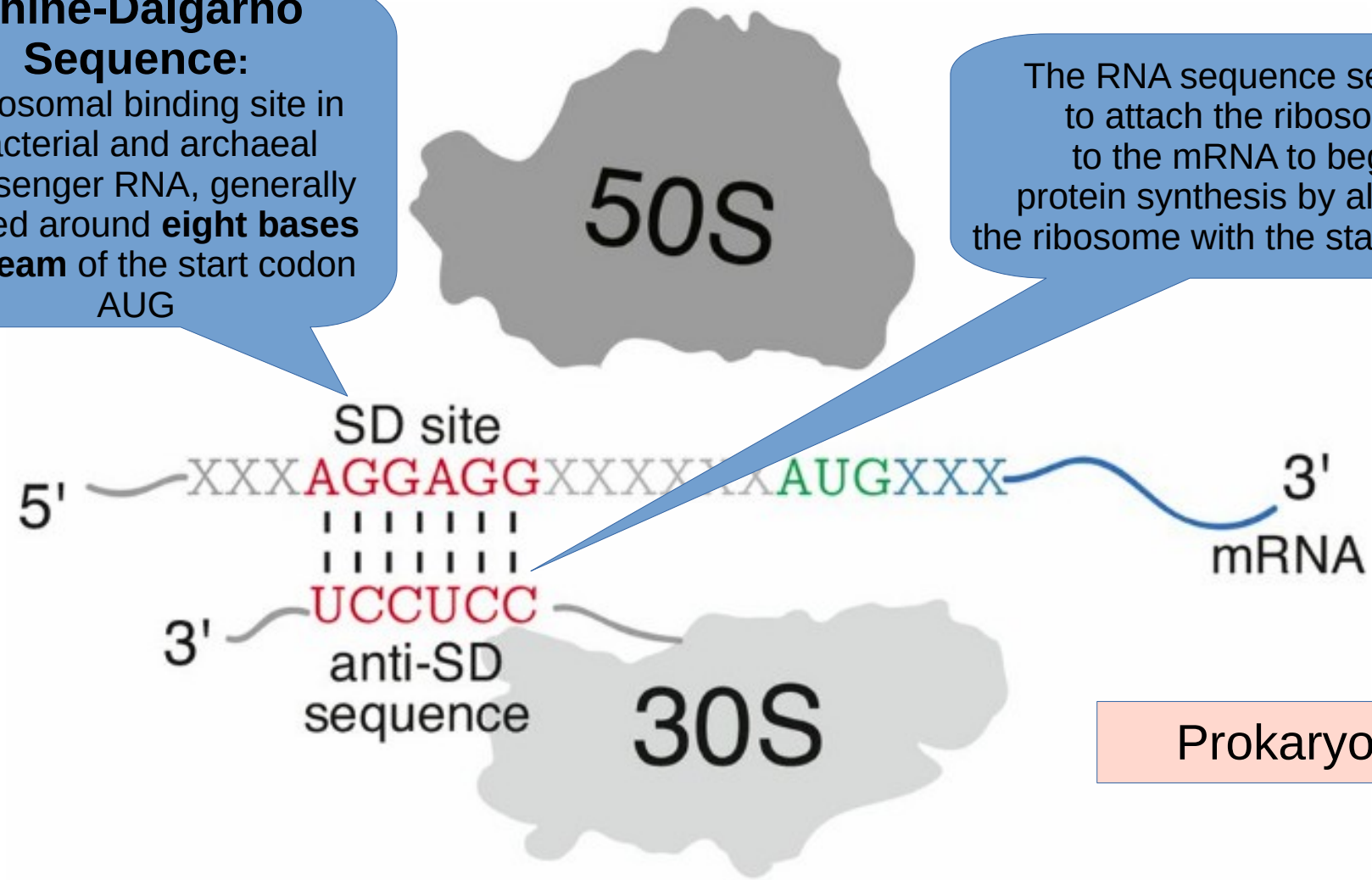
Ribosomal binding site

More details:
https://en.wikipedia.org/wiki/Shine%E2%80%93Dalgarno_sequence

Genetic Land Marks?

Shine-Dalgarno Sequence:
a ribosomal binding site in bacterial and archaeal messenger RNA, generally located around **eight bases upstream** of the start codon AUG

The RNA sequence serves to attach the ribosome to the mRNA to begin protein synthesis by aligning the ribosome with the start codon.

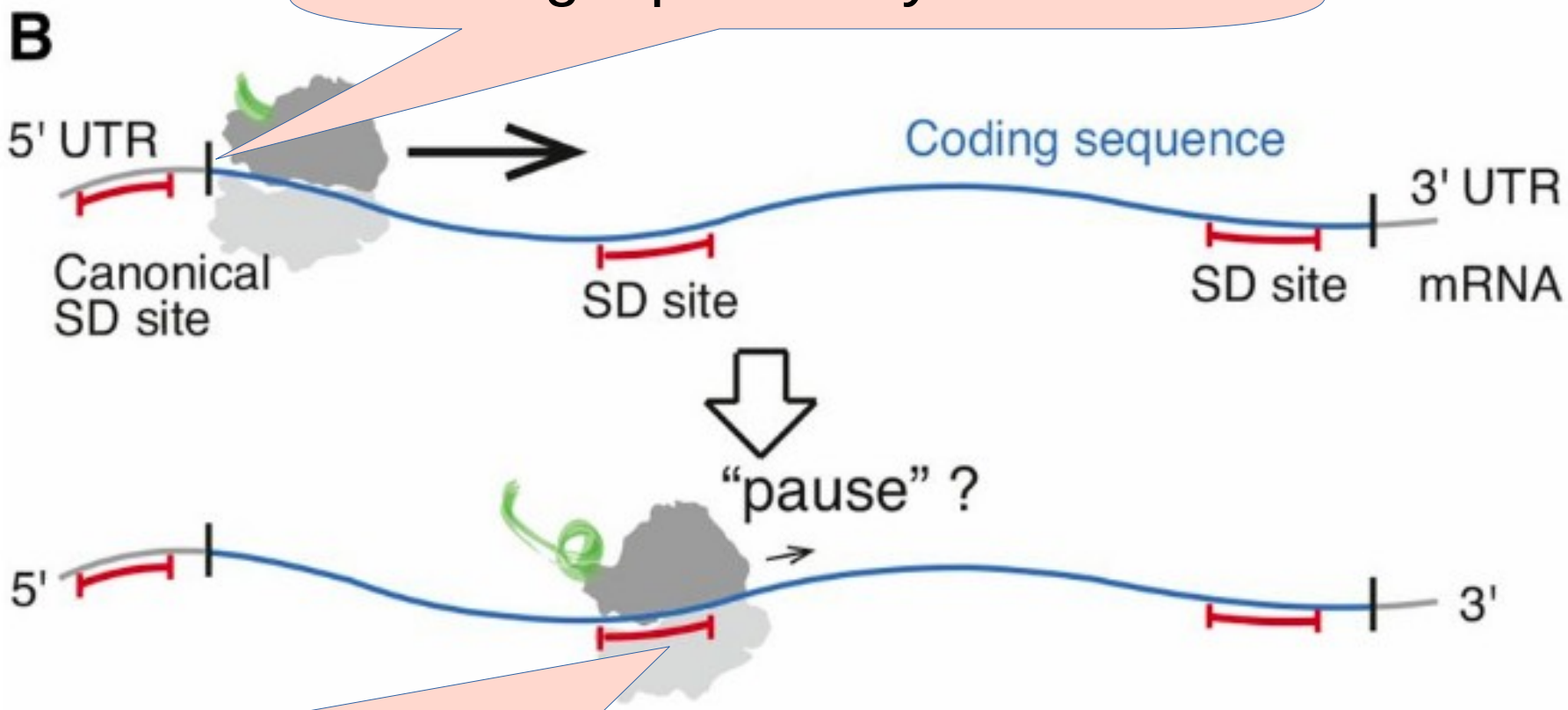


Prokaryotes

Depletion of Shine-Dalgarno Sequences Within Bacterial Coding Regions Is Expression Dependent, Chuyue Yang, Adam J. Hockenberry, Michael C. Jewett and Luís A. N. Amaral, https://www.g3journal.org/content/6/11/3467

Genetic Land Marks?

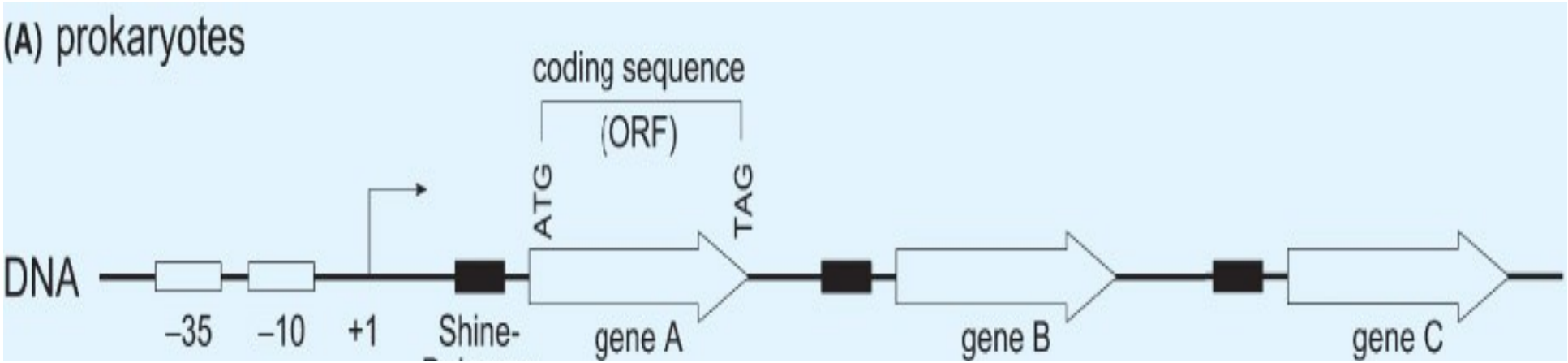
Ribosome attaches here to begin protein synthesis

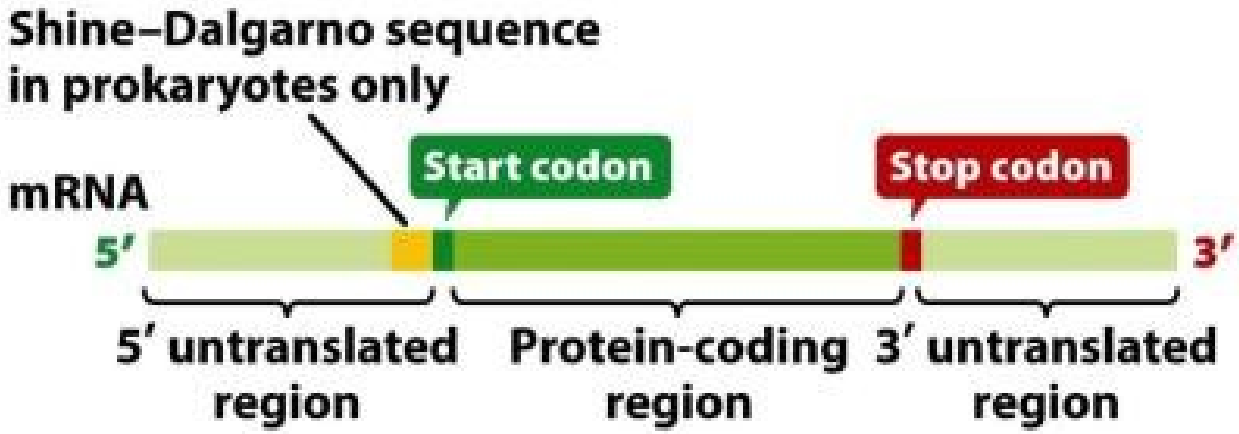


SD sequences within coding sequences may negatively affect translation elongation speed

The Familiar to Find the Unfamiliar

- We look for specific features or *land-marks* in a sequence that may <u>*suggest*</u> that there is a gene at play.
  - The Shine-Dalgarno seq. found upstream of a DNA start codon: ATG





Prediction Algorithms

- Can you find any sense in the below sequence?

Lo gicwi llg etyo ufro mAt oB.
Ima ginat ion wi llge ty
oue ve rywh ere.

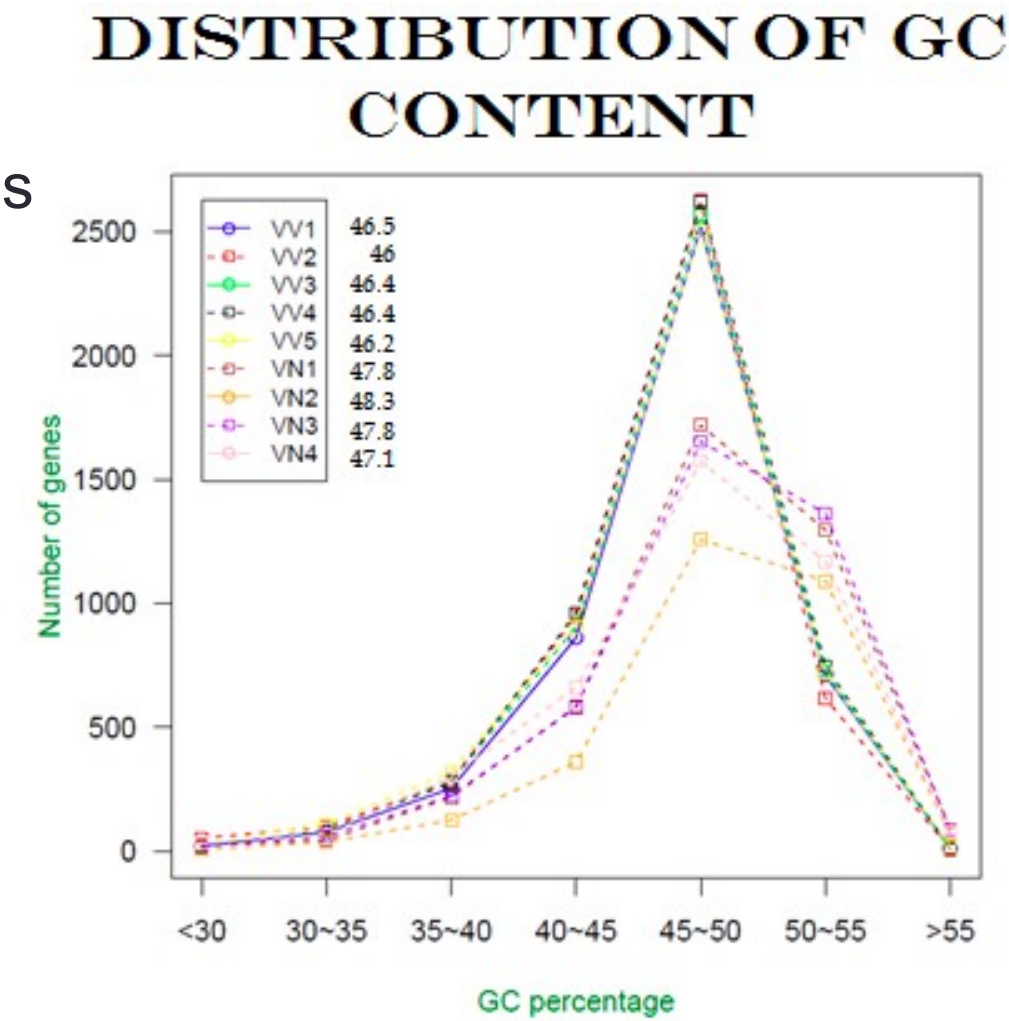
- How did you find the meaning here?
- How would an algorithm do it?

Prediction Algorithms

- Alignment-based – find genes/features based on conserved sequences is well-studied organisms (database searching)
  - Automatic assignment based on sequence similarity (best BLAST hit): gene name, protein name, function
  - Quality vs Quantity: How much time do you have to find this gene? Heuristic-based, or exhaustive search

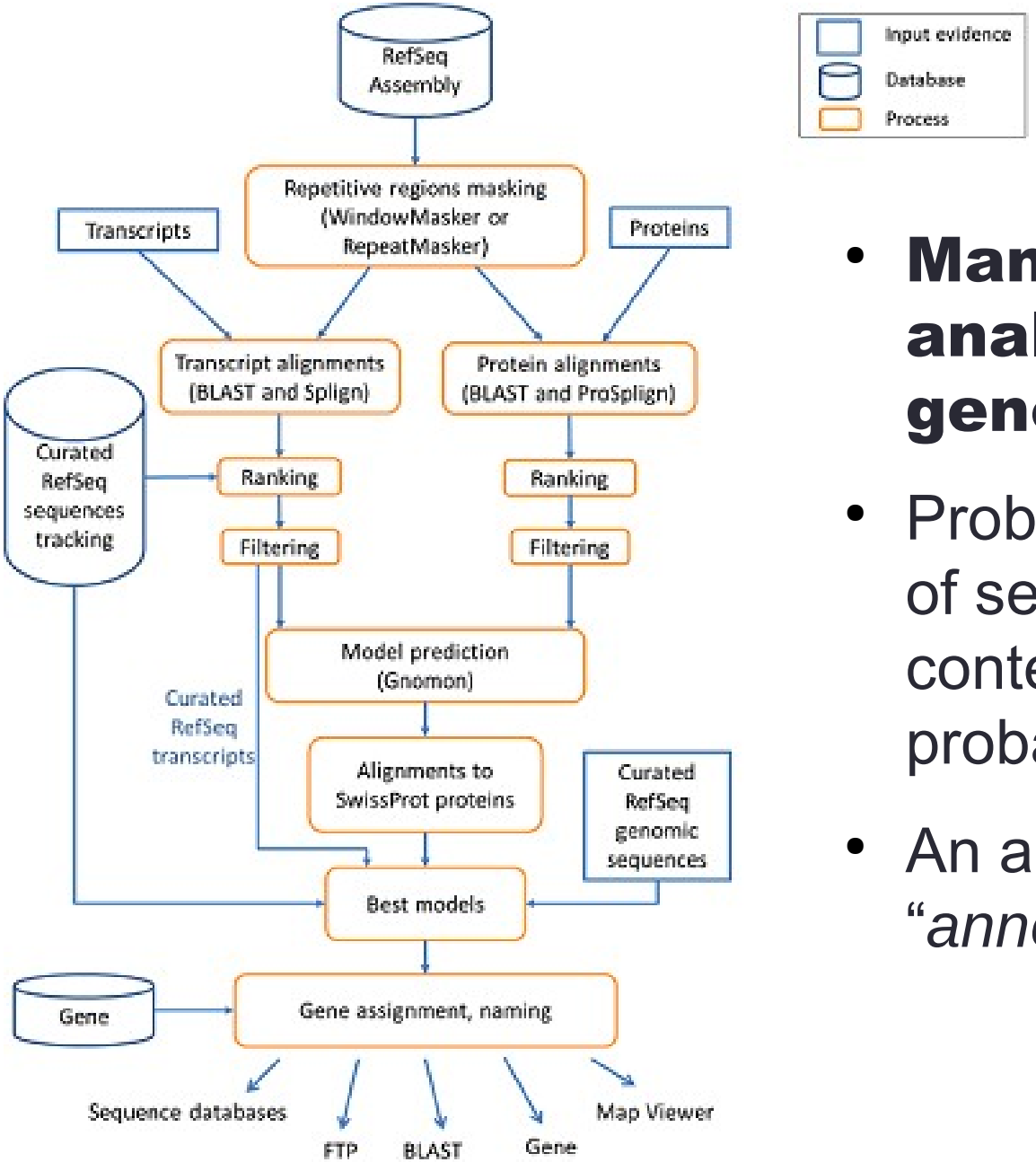
Prediction Algorithms

- Content-based – consider overall properties of the sequence when making predictions
- Nucleotide frequency
- Codon frequency/codon bias



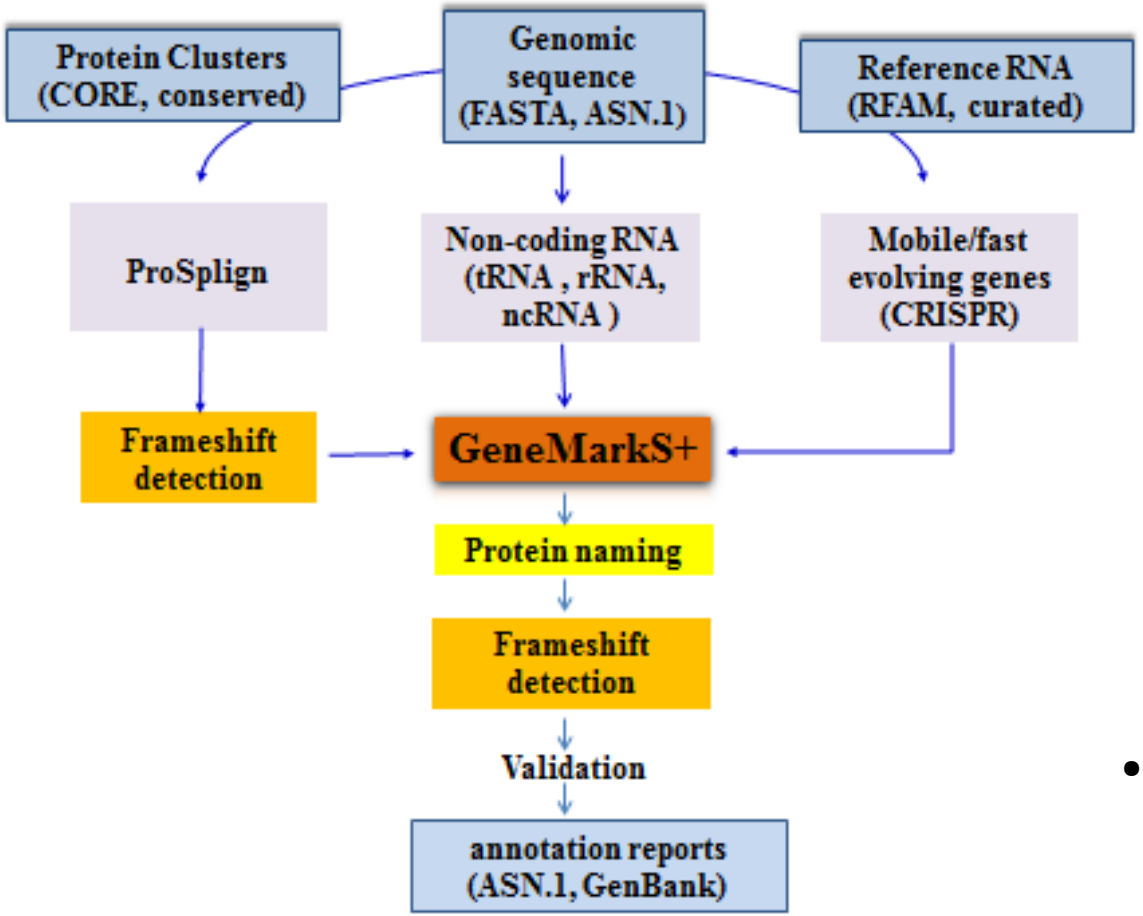
- GC content for all *V. vulnificus* and *V.naverensis* gene predictions (Figure)
- Most of the genomes contained a high percentage of genes with GC contents between 45-50%.

Prediction Algorithms



- **Many stages of analysis to determine genomic artifacts …**
- Probabilistic – combination of sequence-based and content-based plus probability
- An analysis by the "*annotation pipeline*"

NCBI Prokaryotic Annotation Pipeline



- Combines sequence-based algorithm with alignment-based approach
  - Protein-coding genes
  - Structural RNAs (5S, 16S, 23S)
  - Transfer RNAs
  - Small non-coding RNAs
- Rely only on properties of DNA and training set of genes

http://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/

NCBI Prokaryotic Annotation Pipeline

1. Masking
   - Try to identify and ignore non-coding regions
2. Alignment-based predictions
   - *Ask where we have seen this sequence before (BLAST)*
3. Sequence/content-based predictions from alignment-based
4. Best selected (probability), named, and released

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/#assemblies

NCBI Eukaryotic Annotation Pipeline

- The best predictions are selected to describe observed artifacts (purple).
- At the end, the annotation products are formatted and deployed to public resources (yellow).

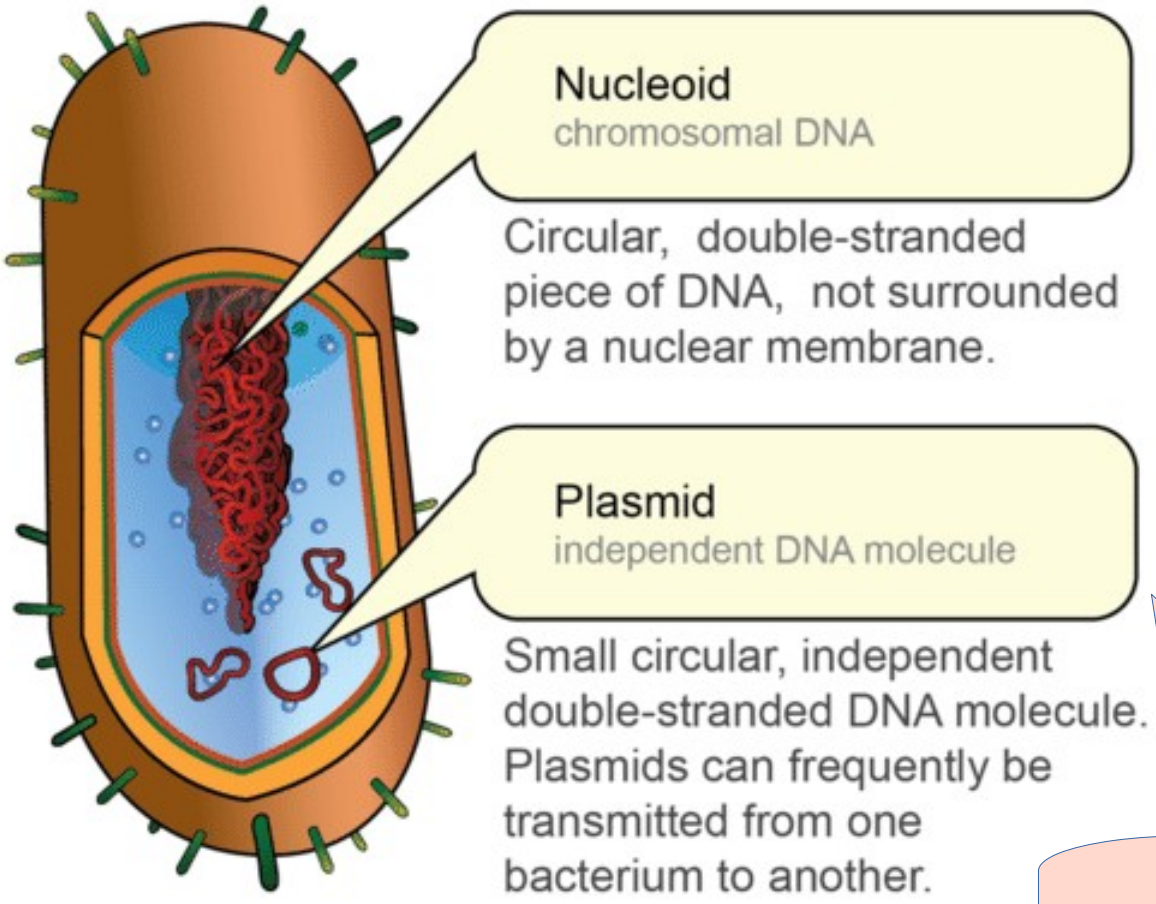
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/#assemblies

Natural Differences

- Algorithms find and compare differences to find genes
- Find similarities to draw conclusions

"Love" in Chinese, Japanese and Korean

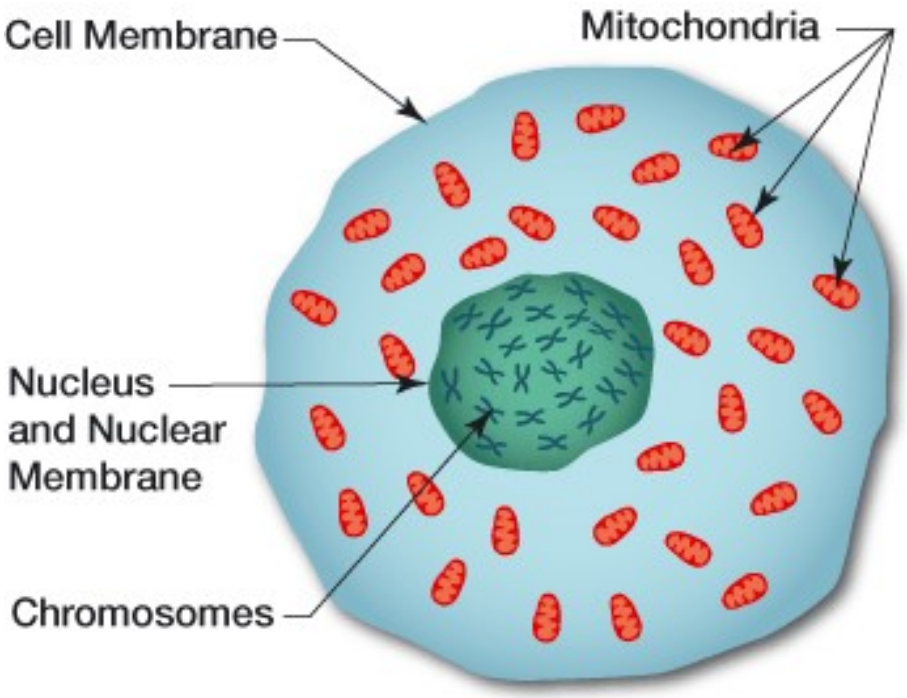
Differences: Prokaryotes



- A circular chromosome
  - "Genome"
- Extra DNA in plasmids
  - smaller, self-replicating

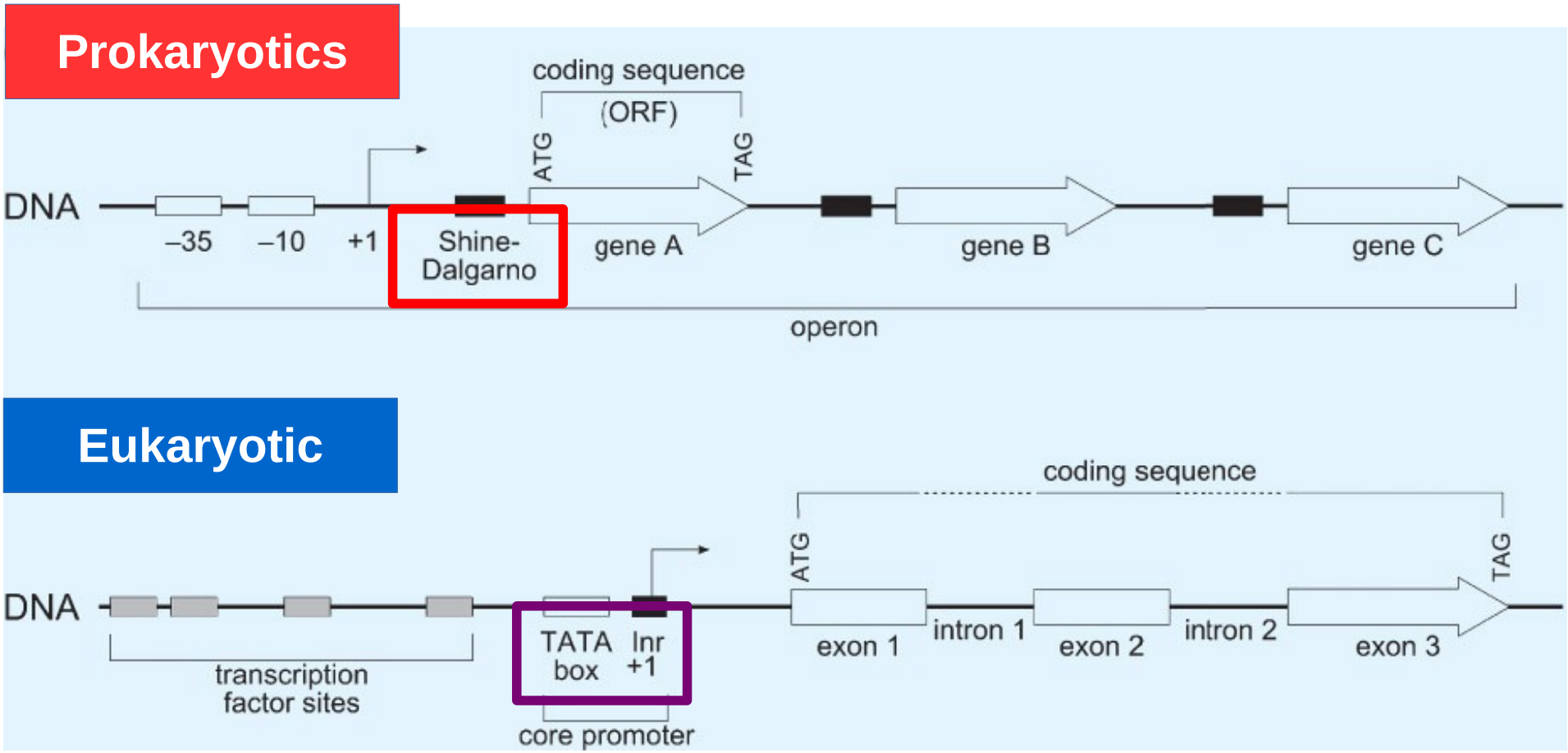
Different types of genomes require different approaches to find genetic differences...

Differences: Eukaryotes



- Multiple linear Chromosomes
  - "Genome"
- Extra DNA in Mitochondria or chloroplast

Types of comparisons



Which DNA is it? Comparison of Landmarks

Prokaryotic versus Eukaryotic Genomes

Organism	Amount of DNA (bp)	# of genes	Genes per million bases
Escherichia coli	4,600,000	4,400	950
Saccharomyces cerevisiae	12,000,000	5,800	480
Drosophila melanogaster	180,000,000	13,700	76
Mus musculus	2,600,000,000	25,000	11
Homo sapiens	2,900,000,000	25,000	10

Eukaryotic cells **Prokaryotic cells**

Consensus Sequences

Table 9.3 Consensus sequences for gene expression in prokaryotes and eukaryotes.

Sequence	Consensus (5′ → 3′)	Function
Prokaryotes		
–10 sequence	TATAAT	RNA polymerase binds to start transcription
–35 sequence	TTGACA 17±2 from –10	RNA polymerase binds to start transcription
Shine-Dalgarno	AGGAGG 5±2 from ATG	Ribosome binds to find start codon
Eukaryotes		
TATA box	TATAWAW	Core promoter; binds TFIID
Inr sequence	YYCARR	Core promoter; contains +1 sequence (C)
GC box	GGGCGG	Transcription factor binding site
CAT box	CAAT	Transcription factor binding site
Kozak consensus	gccRccATGG	Context of start codon
5′ splice site	MAG | GTragt	Bound by spliceosome to remove introns
3′ splice site	cAG | G	Bound by spliceosome to remove introns
intron branch site	CTRAY	3′ end of intron binds to mark for degradation
polyadenylation site	AAUAAA	Cleavage of mRNA for poly(A) tail

Landmarks!

Open Reading Frame (ORF) Vs Coding Sequences (CDS)

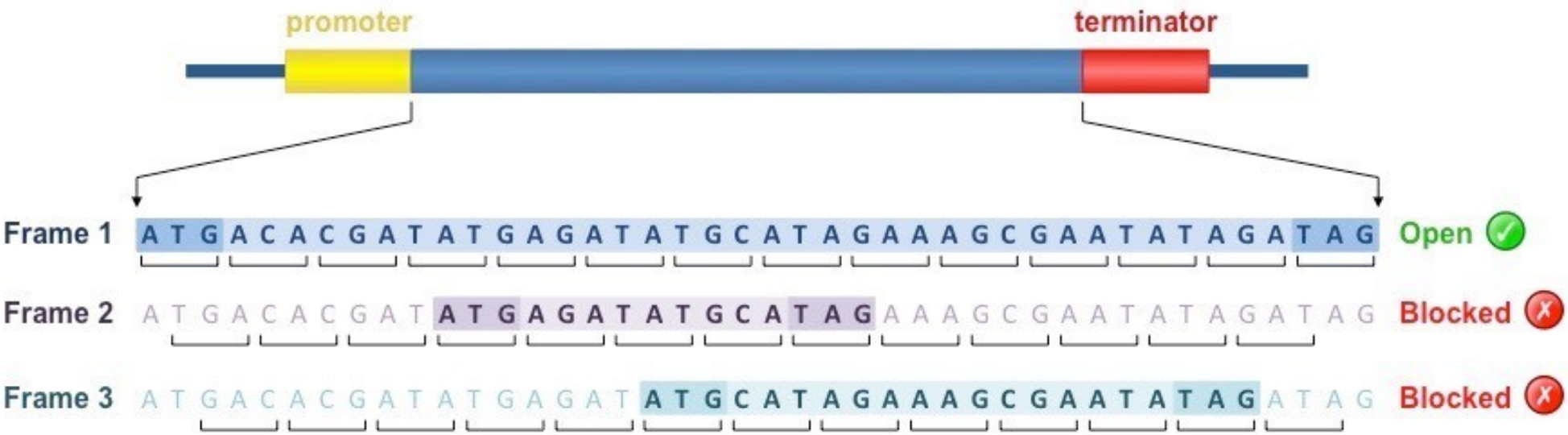
Coding Sequences (CDS) are subsequences (regions in DNA or RNA) of the super-sequence that determine the sequence of amino acids in a protein.

Open Reading Frames (ORF) are series of DNA codons that do not contain STOP codons

Note: All CDS are ORFs, but not all ORFs are CDS.

Open Reading Frame Finders

ORF finders search for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.



Open Reading Frame (ORF) Finding

- Online tools:
  - NCBI:
  - https://www.ncbi.nlm.nih.gov/orffinder/
- Sequence Manipulation Suite:
  - http://www.bioinformatics.org/sms2/orf_find.html

```
 5'                                                    3'
  atgcccaagctgaataagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Bring the Tool!

Up Next!

NCBI's ORF Finder Tool

Quick link:
https://www.ncbi.nlm.nih.gov/orffinder/

Viewing Annotations in NCBI's Open Reading Frame

Cyprinus carpio SIRT1 mRNA, partial cds

GenBank: KF881970.1

FASTA Graphics

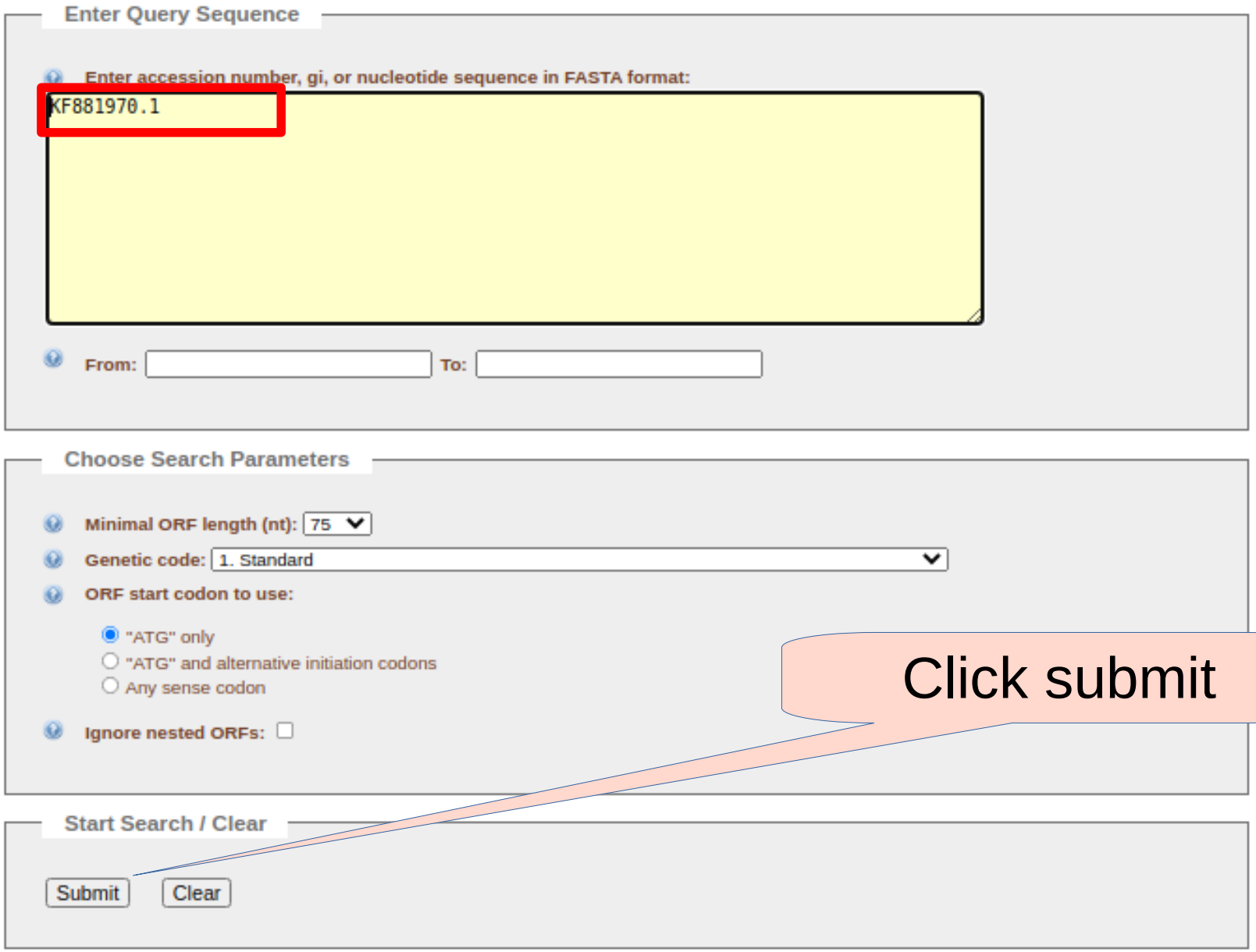
Go to:

```
LOCUS       KF881970                 375 bp    mRNA    linear   VRT 06-NOV-2014
DEFINITION  Cyprinus carpio SIRT1 mRNA, partial cds.
ACCESSION   KF881970
VERSION     KF881970.1
KEYWORDS    .
SOURCE      Cyprinus carpio (common carp)
  ORGANISM  Cyprinus carpio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            Cypriniformes; Cyprinidae; Cyprininae; Cyprinus.
```

SIRT1 Accession

Accession Number:

KF881970.1



Click submit

Results

Cyprinus carpio SIRT1 mRNA, partial cds

ORFs found: 3 Genetic code: 1 Start codon: 'ATG' only

KF881970.1 Find:

(U) ORFfinder_4.19.224388

ORF3

CDS: ORF3
Qualifiers: Partial stop
Name: CDS
Location: complement(2..316)
[Length]
Span on KF881970.1: 315 nt
Protein length: 105 aa
[Positional Info]
KF881970.1 position: 156
CDS position: 161
Protein position: 54
Protein sequence: HLIFILLHGSVEKLW[K]VLAKEDDVWFHDGI

Download FASTA: ORF3

Links & Tools
BLAST nr: KF881970.1 (2..316)
FASTA record: KF881970.1 (2..316)
GenBank record: KF881970.1 (2..316)

ORF3 (104

```
>lcl|ORF3
MRQRLAVDQDL
KLWKVLAKEDD
TGGSR
```

Mark subset... Marked: 0 Download marked set as

Label	Strand	Frame	Start	Stop
ORF3	-	3	316	>2
ORF1	+	2	122	>373
ORF2	-	1	255	82

Protein Locations and Options

Cyprinus carpio SIRT1 mRNA, partial cds

ORFs found: 3 Genetic code: 1 Start codon: 'ATG' only

KF881970.1 Find: Tools

(U) ORFfinder_4.19.224388

ORF3

CDS: ORF3
Qualifiers: Partial stop
Name: CDS
Location: complement(2..316)
[Length]
Span on KF881970.1: 315 nt
Protein length: 105 aa
[Positional Info]
KF881970.1 position: 156
CDS position: 161
Protein position: 54
Protein sequence: HLIFILLHGSVEKLW[K]VLAKEDDVWFHDGI

Download FASTA: ORF3

Links & Tools
BLAST nr: KF881970.1 (2..316)
FASTA record: KF881970.1 (2..316)
GenBank record: KF881970.1 (2..316)

ORF3 (104

>lcl|ORF3
MRQRLAVDQDL
KLWKVLAKEDD
TGGSR

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF3	-	3	316	>2	315 | 104
ORF1	+	2	122	>373	252 | 83
ORF2	-	1	255	82	174 | 57

Get more info about the embedded proteins